December 9, 2024

E-Filed

The Honorable Thomas S. Hixson
United States District Court for the Northern District of California
San Francisco Courthouse, Courtroom E – 15th Floor
450 Golden Gate Avenue
San Francisco, CA 94102

Re: *Kadrey, et al v. Meta Platforms, Inc.*; Case No. 3:23-cv-03417-VC

Dear Magistrate Judge Hixson:

Plaintiffs in the above-captioned action ("Plaintiffs") and Defendant Meta Platforms, Inc. ("Meta") jointly submit this letter brief regarding Meta's search terms in response to Plaintiffs' additional discovery requests. The parties met and conferred on December 2, 2024, but were unable to reach a resolution.

Plaintiffs' Position

It is a bedrock rule of discovery that "parties [must] conduct a reasonable search for documents that are relevant to the claims and defenses." *Raine Grp. v. Reign Capital*, 2022 WL 538336, at *1 (S.D.N.Y. Feb. 22, 2022); *see also Genentech, Inc. v. Trustees of Univ. of Pa.*, 2011 WL 13177617, at *4 (N.D. Cal. Sept. 19, 2011) (finding it reasonable to require a party to conduct a reasonable search to ensure that no relevant or producible documents had been overlooked). Yet Meta's search efforts have been anything but reasonable, forcing Plaintiffs to burden the Court yet again with avoidable motion practice. Meta's redoubled efforts to hide the ball in response to Plaintiffs' timely served "additional discovery" should neither be countenanced, nor rewarded.

I. Meta's Search Terms Do Not Adequately Address Fundamental Issues.

The search terms do not connect copyright to infringement or piracy. Plaintiffs' current claim is for copyright infringement, and Meta's produced documents show Meta employees referred to the books in LibGen that Meta downloaded for use with the Llama models as "pirated works." *See, e.g.*, Meta_Kadrey_74729-31 at 74730 (Eleanora Presani tells Xavier Martinet that LibGen is "pirated material," and that both LibGen and SciHub are "illegal pirated websites"). Meta's search terms lack any variations of the words "infringement," "piracy," "violation," "illegal," "liability," or "torrent."[1] *See* Ex. B (showing the search terms Meta ran). Instead, Meta searched for *positive* words—compensation, consent, licensing, credit, attribution—far less likely to capture documents regarding Meta's concerns about stealing or pirating copyrighted material rather than licensing it. Plaintiffs request that the Court order Meta to incorporate Plaintiffs' proposed terms into all relevant searches in response to Plaintiffs' "additional discovery" requests. *See* Ex. A for Plaintiffs' suggested search terms and the relevant RFPs corresponding to each search term.

The search terms do not address removal of copyright management information. Meta's terms are also deficient because they do not include terms like "delete," "remove," "strip," "cut," "strike," or "excise," and do not incorporate the common acronym "CMI" (copyright management information). Such terms capture documents evidencing Meta's illegal removal of CMI from copyrighted works in LibGen and other datasets used to train Llama models. Any such removal has clear relevance to, *inter alia*, Meta's intentionality and willfulness and the scope and size of the putative class. *See* Ex. A, Term 2.

The search terms dramatically limit results concerning Meta's fair use defense. Meta previewed in discovery that it **may** attempt to argue its "legal review" of copyright issues is relevant to Plaintiffs' allegations of willfulness, while claiming privilege over that review yet disclaiming the advice of counsel defense. Yet Meta searched for "fair use" ***only with the limitation that it be mentioned within 20 words of the named Plaintiffs***. This is plainly inadequate. At a minimum, Meta must search for and produce all documents in which "fair use" is mentioned with respect to its Llama models, including datasets used or considered to be used in training Llama. Plaintiffs request that the Court order Meta to expand its search terms to locate those essential documents in connection with Plaintiffs' additional discovery. *See* Ex. A, Term 3.

Meta's search terms are lacking with respect to publishers and datasets. Meta's terms do not include variations of the word "publisher," let alone the names of the numerous publishers Meta

[1] Meta's counsel confirmed in a meet and confer that Meta had re-checked documents previously located with its existing search terms for responsiveness to Plaintiff's Fifth RFPs, *i.e.*, the set to which Meta responded after extension of the discovery deadline.

initially contacted to negotiate licenses. Also, the terms "the pile," "sci-hub" and "the eye" have improper limitations applied such as appearance with the terms "dataset" or "training data." Plaintiffs request that the Court order Meta to add generic terms for book publishers and search for the indicated datasets without limitation. Ex. A, Terms 4-5.

II. Meta Should Search Email and Workplace Chat Without Custodian Restrictions.

Meta has refused to fully search company email and Workplace[2] for documents responsive to Plaintiffs' additional discovery. Meta disputes these sources are non-custodial yet confirmed that they can be accessed and searched "on the back end" by Meta's internal "discovery team." *See also* Dkt. No. 267 at 6 nn. 5 & 6. In other words, they are centrally searchable and thus non-custodial, as this Court's ESI Discovery Checklist explains, *see* § III (discovery must be prioritized from systems like email).[3] Instead, Meta has searched only the email and Workplace files for 15 custodians even though it can access remotely its many other AI employees' email and Workplace files. Plaintiffs request the Court order Meta to apply its search terms through its company email and Workplace for all AI employees (and former employees), not just 15 of them.

III. Meta Searched WhatsApp for Just Three Custodians, and, in so Doing, Employed Insufficient Search Terms.

Meta searched WhatsApp for only Mark Zuckerberg and two other custodians. Meta has represented it merely asked whether custodians used WhatsApp for work, *see* Dkt. No. 267 at 5-6, without any independent verification by Meta's counsel. This is improper. *See, e.g., DR Distribs. v. 21 Century Smoking*, 513 F. Supp. 3d 839, 935 (N.D. Ill. 2021) (attorneys must make "reasonable inquiry to verify that the client accurately captured [the] universe" of relevant documents). Plaintiffs respectfully request the Court order Meta to search all of its custodians' WhatsApp accounts for responsive communications. Additionally, Meta searched just 11 terms in the three WhatsApp accounts it did search, but excluded obviously relevant terms like "llm," "language model," "dataset," "ChatGPT," "OpenAI," "infringe!," and "pirat!" Plaintiffs respectfully request the Court order Meta to employ those search terms in WhatsApp.

IV. Meta Must Perform Searches Covering the Entire Class Period.

Meta unilaterally and improperly limited the relevant period for its productions to January 1, 2022, to the present. Unlike Plaintiffs' challenge to Meta's use of this circumscribed date range in connection with Plaintiffs' early discovery requests, which this Court denied as untimely, *see* Dkt. No. 288, this challenge concerns Meta's imposition of that limited date range to Plaintiffs' timely served additional discovery. Further, Meta has repeatedly asserted that that there simply is nothing to discover before January 1, 2022, rendering any burden on Meta de minimis. Plaintiffs thus request that the Court order Meta to expand the time period of its searches in response to Plaintiffs' additional discovery to the beginning of the class period.

[2] Workplace is a company platform that combines chat, video, and groups, among other things.

[3] *See, e.g.*, The Sedona Conference, The Sedona Principles, 19 Sedona Conf. J. 1, 103 (2018) ("sources of non-custodial relevant information" include "structured systems and databases, and other non-custodial sources such as collaboration tools, social media," and any other "shared areas (such as public folders, discussion databases, and shared network folders) that are not regarded as belonging to any specific employee").

Meta's Position

I. Meta's Search Terms Were Tied to Plaintiffs RFPs and Plaintiffs' Did Not Follow The ESI Order Process To Seek Additional Search Terms. Plaintiffs' complaints about Meta's search terms are disconnected from and seek to rewrite Plaintiffs' ***136*** actual RFPs in this case. Plaintiffs now seek to sweep in issues that are not covered by the RFPs and add matters related to their belated motion to amend the complaint (ECF 300, to which Meta will separately respond). Meta's carefully tailored search terms are responsive to Plaintiffs' actual RFPs, as written. As the Court can see in Plaintiffs' Ex. B (reproducing Meta's search term and hit count disclosure), Meta conducted broad searches, including searching dataset and product names (e.g., Llama) without any qualifiers. Ex. B at 9.

Plaintiffs' request flouts the Court-ordered process for requesting additional search terms. The ESI Order dictates that the ***producing party*** selects the search terms and then discloses them. ECF 101 at 3, 4. If Plaintiffs object to the sufficiency of Meta's search terms, the ESI Order dictates the process that should follow, which involves requesting that Meta review an additional null set sample, and if relevant responsive documents are identified, iterating on additional search terms specifically targeting those documents – a process that would itself not be feasible given the limited time remaining in discovery. *Id.* at 4-5. Plaintiffs ignored this Court-ordered process and instead– less than two weeks before the close of discovery, demanded that their own search terms be run instead. Plaintiffs' proposal here contravenes the ESI Order and should be rejected.

"Infringement" and "Piracy"-Type Search Terms Are Not Called for by the Cited RFPs. The cited RFPs do not mention anything about "infringement," "piracy," "violation," "illegal," "liability," or "torrent." RFP Nos. 77 and 105 are about licensing. And RFP No. 106 is about guidelines for using copyrighted materials in the models. Plaintiffs do not explain how any of their proposed search terms link to any of the cited RFPs. Plaintiffs' exemplary citations of Meta-produced documents demonstrate that Meta's existing searches yielded the types of documents they seek.

Alleged Removal of CMI is Not Relevant to Current Claims or Called for By Cited RFPs. Again, Plaintiffs' cited RFPs do not mention anything about copyright management information (CMI). And understandably so, as Plaintiffs' related DMCA claim was dismissed from the case over a year ago, *see* ECF 56 at 3, rendering this discovery irrelevant. If Plaintiffs now seek CMI-related discovery in light of their pending motion to amend, that is improper and contradicts Plaintiffs' representation that they do not require any "additional discovery beyond what is needed for the copyright infringement claim." ECF 301 at 1. In any case, Plaintiffs' motion to amend cited documents produced months ago regarding alleged removal of CMI, belying their claim that Meta's existing search terms did not capture such materials. ECF 301-7.

The Cited RFPs Do Not Support the Requested Searches on Fair Use. Of the RFPs cited by Plaintiffs, only one (No. 123) mentions fair use. And that is a narrow RFP regarding ***communications with third parties*** about Plaintiffs' allegations and Meta's defenses. None of the Plaintiffs' RFPs request what Plaintiffs now seek, *i.e.*, "all documents in which 'fair use' is mentioned with respect to its Llama models, including datasets used or considered to be used in training Llama." *Supra* at 1. Regardless, Plaintiffs do not dispute that Meta has produced many documents about fair use as it relates to the Llama models. Indeed, as shown in Ex. B, Meta conducted broad searches for dataset names and "Llama." As to the other RFPs cited by Plaintiffs, No. 117 was rejected by the Court last week (ECF 315 at 9); No. 81 seeks documents about "Shadow Dataset"[4] training, not fair use; No. 77 is about "coherent storytelling," with no articulated connection to fair use; and No. 130 is plainly about licensing agreements, not fair use.

Meta Searched for Publishers and No Explanation is Provided for the Datasets. Plaintiffs falsely assert that Meta failed to use publisher names as search terms in response to RFPs 77 and 130, when in fact it did. Ex. B at 2 and 9 (search terms starting with [redacted]). Plaintiffs do not explain how those are not adequate. Regarding datasets, Plaintiffs cite only one RFP mentioning any of the "dataset" terms they now seek to use–RFP 83, which pertains to the role "The Eye" had in distributing the Books3 dataset, and Meta conducted this search. Ex. B at 6, 11. Plaintiffs do not and cannot identify any RFP covering their other proposed terms, such as "The Pile" or "Sci-Hub."

II. Email and Workplace Chat Are Custodial Sources That Were Searched for the 15 Custodians. Meta's Email and Workplace Chat are quintessential custodial data sources: documents in these systems are created, identified, stored, and can be searched by reference to individual Meta employees.[5] Meta has searched for and produced Email and Workplace Chat for all 15 Meta custodians, as was its obligation. The Sedona Principles cited by Plaintiffs do not support characterizing email or instant messaging as "non-custodial" data sources (which in the excerpt are large "structured systems and databases" (e.g. a SQL Database) and other shared repositories like "shared network folders" not managed by or easily narrowed by custodian), and Plaintiffs' erroneous interpretation would upend common discovery practice, as in 2024, many custodial sources have centralized access by IT.

Notably, the Court previously rejected Plaintiffs' attempted end-run around the ESI Order by seeking to characterize custodial data sources as non-custodial. *See* ECF 279 at 3-4. Requiring Meta to search email and chat messages for "all AI employees (and former employees)" would contravene the Orders of this Court, which authorized only 15 Meta custodians. *See* ECF 101 (authorizing 10); ECF 212 (allowing 5 more). To be clear, Meta has collected and produced from custodial and non-custodial sources when identified during Meta's reasonable search, including cases where no ESI custodian was involved.

III. Meta Produced Relevant WhatsApp Messages. WhatsApp is also a custodial data source, and Meta conducted a reasonable search for relevant WhatsApp messages, including searching information from all custodians that represented they may have relevant messages, which was confirmed by Plaintiffs' deposition questioning. Plaintiffs fail to provide any justification to compel Meta to search other custodians who did not use WhatsApp for work. With respect to

[4] The Court denied Plaintiffs' prior motion to compel regarding the definition of Shadow Datasets which required Meta to speculate what datasets are considered "shadow." ECF 315 at 7-8.

[5] Plaintiff interchangeably uses "Workplace" and "Workplace Chat" (also known as "Workchat"). Workplace is a collaboration platform where teams and individuals can post information (similar to a Facebook page) and was a resource that Meta collected materials from in discovery. Workplace Chat is a distinct instant messaging platform that individuals at Meta use to communicate with one another, and is a custodial ESI source like email.

search terms, Plaintiffs' request violates the ESI Order (*see supra*) and fails to connect any of the requested search terms to any specific RFPs.

IV. The Court Already Rejected Plaintiffs Class Period Date Range. The Court previously rejected Plaintiffs' request to extend the Jan. 1, 2022 start date for document collection and production. ECF 288 at 3. Meta disclosed its search date range to Plaintiffs in February 2024, and this date was thoughtfully selected because it is many months before Meta began developing Llama in the Fall of 2022. There is thus no basis to go back in time further, and doing so in the few days before the close of discovery would also impose massive, disproportionate, and impossible burdens on Meta–precisely the type of "train wreck," ECF 279 at 5, that the Court has refused to allow. Specifically, requiring Meta to use a new, unjustified start date for document collection would require the wholesale recollection of materials before searching, review, and production, and would likely take months to complete. Meta should deny this unjustified request, as it has previously. *See* ECF 288 at 3.

Plaintiffs' Reply

ESI Order/Timing. Plaintiffs repeatedly asked Meta to identify its search terms no later than 10/9. *E.g.,* Dkt. 247 at 25. Meta cites the ESI Order, but flouted its procedure, which "dictates that the producing party selects the search terms and then discloses them." ECF 101. Meta only disclosed its terms ***six weeks*** after Plaintiffs requested them, on 11/18 (with several follow-ups to which Meta said its terms were forthcoming). Meta should not be rewarded for playing these kinds of games. And if Meta needs more time to comply, it can seek time from Judge Chhabria.

Piracy/Infringement Search Terms. RFP 105 is not just about licensing; it is a request for documents concerning training data, "including but not limited to the inclusion of any copyrighted material within data used to train Llama Models." Leaving out "piracy," "infringement," "torrent," and the like only conceals evidence on the willful use of such material.

Removal of CMI Search Terms. Just because requested search terms are also relevant to the proposed amended complaint does not make them *irrelevant* to the existing infringement claim. The stripping of CMI is clearly relevant to Meta's willfulness in infringing Plaintiffs' copyrights.

Fair Use. Meta ignores the plain text of RFPs concerning Meta's decision-making on using copyrighted data. *See, e.g.*, RFP 81 "related to [Meta's] decision to use Shadow Datasets for training." Meta itself raises fair use in that context. *See, e.g.*, Meta's Answer (Dkt. 154) at 11.

Email/Work Chat. Meta argues running its search terms through Meta's email and Workplace servers (at least for its AI employees) "would upend common discovery practice." That's quite the position (for which Meta cites nothing) given this Court's ESI guidance.[6] Further, the ESI Order nowhere allows Meta to restrict its use of search terms to 15 custodians' emails, which be an absurdly limited search for a company of Meta's size. *See* Dkt. 101, § 7; *see also Crocs, Inc. v. Effervescent, Inc.*, 2017 WL 3891697, at *2 (D. Colo. Feb. 3, 2017) ("the company is the custodian of all emails in all accounts associated with the company"). The Court should order Meta to run its search terms through its AI employees' work email and produce all responsive, non-privileged documents, and if it needs more time then the parties can go to Judge Chhabria.

[6] https://www.cand.uscourts.gov/wp-content/uploads/court-programs/cja/electronic-discovery/Recommended-E-Discovery-Practices.pdf (an "email database," like the one Meta controls and its discovery team can search remotely on the back end, is "typically part of a large database of linked, organized, and searchable records").

By: /s/ *Bobby Ghajar*

Bobby A. Ghajar
Colette A. Ghazarian
COOLEY LLP
1333 2nd Street, Suite 400
Santa Monica, CA 90401
Telephone: (310) 883-6400
Facsimile: (310) 883-6500
Email: bghajar@cooley.com
cghazarian@cooley.com

Mark R. Weinstein
Elizabeth L. Stameshkin
COOLEY LLP
3175 Hanover Street
Palo Alto, CA 94304
Telephone: (650) 843-5000
Facsimile: (650) 849-7400
Email: mweinstein@cooley.com
lstameshkin@cooley.com

Kathleen R. Hartnett
Judd D. Lauter
COOLEY LLP
3 Embarcadero Center, 20th Floor
San Francisco, CA 94111
Telephone: (415) 693-2071
Facsimile: (415) 693-2222
Email: khartnett@cooley.com
jlauter@cooley.com

Phillip Morton
COOLEY LLP
1299 Pennsylvania Avenue, NW, Suite 700
Washington, DC 20004
Telephone: (202) 842-7800
Facsimile: (202) 842-7899
Email: pmorton@cooley.com

Angela L. Dunning
CLEARY GOTTLIEB STEEN & HAMILTON LLP
1841 Page Mill Road, Suite 250
Palo Alto, CA 94304
Telephone: (650) 815-4121

By: /s/ *Maxwell V. Pritt*

BOIES SCHILLER FLEXNER LLP
David Boies (*pro hac vice*)
333 Main Street
Armonk, NY 10504
(914) 749-8200
dboies@bsfllp.com

Maxwell V. Pritt (SBN 253155)
Joshua M. Stein (SBN 298856)
44 Montgomery Street, 41st Floor
San Francisco, CA 94104
(415) 293-6800
mpritt@bsfllp.com
jstein@bsfllp.com

Jesse Panuccio (*pro hac vice*)
1401 New York Ave, NW
Washington, DC 20005
(202) 237-2727
jpanuccio@bsfllp.com

Joshua I. Schiller (SBN 330653)
David L. Simons (*pro hac vice*)
55 Hudson Yards, 20th Floor
New York, NY 10001
(914) 749-8200
jischiller@bsfllp.com
dsimons@bsfllp.com

Interim Lead Counsel for Plaintiffs

Facsimile: (650) 849-7400
Email: adunning@cgsh.com

Attorneys for Defendant Meta Platforms, Inc.

ATTESTATION PURSUANT TO CIVIL LOCAL RULE 5-1(h)

I hereby attest that I obtained concurrence in the filing of this document from each of the other signatories. I declare under penalty of perjury that the foregoing is true and correct.

Dated: December 9, 2024

BOIES SCHILLER FLEXNER LLP

/s/ Maxwell V. Pritt
Maxwell V. Pritt
Reed Forbush
Jay Schuffenhauer

Attorneys for Plaintiffs